

Knowledge Discovery and Data Mining

Unit # 6

Terminology

- **True Positive**: The number of positive examples **correctly predicted** by the classification model.
- **False Negative**: The number of positive examples **wrongly predicted** as negative by the classification model.
- **False Positive**: The number of negative examples **wrongly predicted** as positive by the classification model.
- **True Negative**: The number of negative examples **correctly predicted** by the classification model.

Terminology (Cont'd)

- The **true positive rate (TPR)** or **sensitivity** is defined as $TPR = TP / (TP + FN)$.
- The **true negative rate (TNR)** or **specificity** is defined as $TNR = TN / (TN + FP)$.
- The **false positive rate (FPR)** is defined as $FPR = FP / (TN + FP)$.
- The **false negative rate (FNR)** is defined as $FNR = FN / (TP + FN)$.

ROC (Receiver Operating Characteristic)

- Developed in 1950s for signal detection theory to analyze noisy signals
 - Characterize the trade-off between positive hits and false alarms
- ROC curve plots TPR (on the y-axis) against FPR (on the x-axis)
- Remember that TPR represents “sensitivity” while FPR represents “100 – specificity”.

How to Construct an ROC curve

Instance	P(+ A)	True Class
1	0.95	+
2	0.93	+
3	0.87	-
4	0.85	-
5	0.85	-
6	0.85	+
7	0.76	-
8	0.53	+
9	0.43	-
10	0.25	+

- Use classifier that produces posterior probability for each test instance $P(+|A)$
- Sort the instances according to $P(+|A)$ in decreasing order
- Apply threshold at each unique value of $P(+|A)$
- Count the number of TP, FP, TN, FN at each threshold
- TP rate, $TPR = TP/(TP+FN)$
- FP rate, $FPR = FP/(FP + TN)$

Sajjad Haider

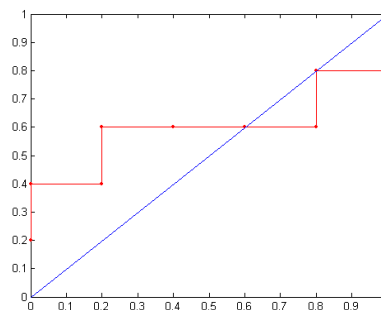
Fall 2012

5

How to construct an ROC curve

Class	+	-	+	-	-	-	+	-	+	+	
Threshold	0.25	0.43	0.53	0.76	0.85	0.85	0.85	0.87	0.93	0.95	1.00
TP	5	4	4	3	3	3	3	2	2	1	0
FP	5	5	4	4	3	2	1	1	0	0	0
TN	0	0	1	1	2	3	4	4	5	5	5
FN	0	1	1	2	2	2	2	3	3	4	5
TPR	1	0.8	0.8	0.6	0.6	0.6	0.6	0.4	0.4	0.2	0
FPR	1	1	0.8	0.8	0.6	0.4	0.2	0.2	0	0	0

ROC Curve:



Sajjad Haider

Fall 2012

6

Lift and Gain Charts

- Very commonly used in the marketing research.
- **Lift** is a measure of the effectiveness of a predictive model calculated as the ratio between the results obtained with and without the predictive model.
- A lift chart consists of a lift curve and a baseline
- The greater the area between the lift curve and the baseline, the better the model

Sajjad Haider

Fall 2012

7

Example

http://www2.cs.uregina.ca/~dbd/cs831/notes/lift_chart/lift_chart.html

- Using the response model $P(x)=100-AGE(x)$ for customer x and the data table, construct the cumulative gains and lift charts. Ties in ranking should be arbitrarily broken by assigning a higher rank to who appears first in the table.

<i>Customer Name</i>	<i>Height</i>	<i>Age</i>	<i>Actual Response</i>
Alan	70	39	N
Bob	72	21	Y
Jessica	65	25	Y
Elizabeth	62	30	Y
Hilary	67	19	Y
Fred	69	48	N
Alex	65	12	Y
Margot	63	51	N
Sean	71	65	Y
Chris	73	42	N
Philip	75	20	Y
Catherine	70	23	N
Amy	69	13	N
Erin	68	35	Y
Trent	72	55	N
Preston	68	25	N
John	64	76	N
Nancy	64	24	Y
Kim	72	31	N
Laura	62	29	Y

Sajjad Haider

Fall 2012

Example: Steps 1 & 2

1. Calculate $P(x)$ for each person x
2. Order the people according to rank $P(x)$

Customer Name	$P(x)$	Actual Response
Alex	88	Y
Amy	87	N
Hilary	81	Y
Philip	80	Y
Bob	79	Y
Catherine	77	N
Nancy	76	Y
Jessica	75	Y
Preston	75	N
Laura	71	Y
Elizabeth	70	Y
Kim	69	N
Erin	65	Y
Alan	61	N
Chris	58	N
Fred	52	N
Margot	49	N
Trent	45	N
Sean	35	Y
John	24	N

Sajjad Haider

Fall 2012

9

Example: Step 3

- Calculate the percentage of total responses for each cutoff point
 - Response Rate = Number of Responses / Total Number of Responses (10)

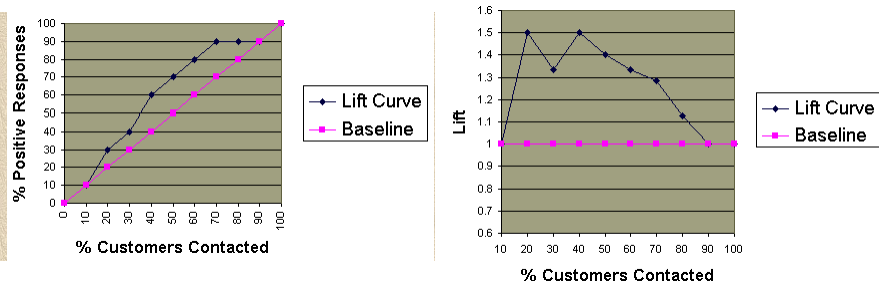
Total Customers Contacted	Number of Responses	Response Rate
2	1	10%
4	3	30%
6	4	40%
8	6	60%
10	7	70%
12	8	80%
14	9	90%
16	9	90%
18	9	90%
20	10	100%

Sajjad Haider

Fall 2012

10

Example: Gains and Lift Charts



Sajjad Haider

Fall 2012

11

Exercise

- Draw gains and lift charts.

Instance	$P(+ A)$	True Class
1	0.95	+
2	0.93	+
3	0.87	-
4	0.85	-
5	0.85	-
6	0.85	+
7	0.76	-
8	0.53	+
9	0.43	-
10	0.25	+

Sajjad Haider

Fall 2012

12

Practice Exercises

- Draw ROC for data set given in slide 9.
- Apply supervised and unsupervised discretization techniques on “Attribute 2” of data set given on slide 18 of Unit # 4.
- Form decision trees using Entropy, Gini and Gain_ratio as splitting criteria on the above data set (after discretization).